Physical Communication 61 (2023) 102218

Contents lists available at ScienceDirect

Physical Communication

journal homepage: www.elsevier.com/locate/phycom

Full length article

Energy-efficient underwater acoustic communication based on Dyna-Q with an adaptive action space[☆]

Cheng Fan [*], Zhaohui Wang, Kaichen Yang

Department of Electrical and Computer Engineering, Michigan Technological University, United States of America

ARTICLE INFO

Keywords:
Underwater acoustic communications
Reinforcement learning
Dyna-Q
Adaptive action space

ABSTRACT

The underwater acoustic (UWA) channel exhibits large spatial–temporal dynamics. This work focuses on adapting the transmission power to the channel condition to achieve energy efficiency in UWA communications. To tackle the unknown channel variation, a reinforcement learning (RL) algorithm called Dyna-Q is introduced. Consider the continuous variation of UWA channels. Instead of using a fixed action space in the Dyna-Q, an adaptive action space with an updated Q-value iteration is proposed in this work. The switching of the transmission power level occurs on a block-by-block basis during transmission, aiming to maximize the system's long-term energy efficiency performance. Using the widely-used communication technique, Orthogonal Frequency-Division Multiplexing (OFDM), as an example, both simulations and experimental data processing results demonstrate that the proposed method outperforms two comparative methods, including the original Dyna-Q with fixed action spaces.

1. Introduction

The underwater acoustic (UWA) channel is a complex and unpredictable process characterized by significant spatial and temporal variations. To ensure efficient UWA communications, it is necessary to adapt the communication strategy to the ever-changing channel conditions. This study investigates the use of reinforcement learning (RL) to enable adaptive switching among different transmission power levels based on the current channel state.

Due to the widespread application of batteries in underwater communication nodes, energy efficiency (EE) plays a crucial role in UWA communication systems, aiming to regulate the transmission power of acoustic devices to enhance overall communication performance under the constraint of limited energy. The energy-efficiency maximization problem can be formulated as a Markov decision process (MDP), where the transmitter's action involves selecting an appropriate power level, and the system's state depends on both the transmitted power and the channel condition. RL is a widely-used learning approach to solve MDP problems, in which an agent takes intelligent actions in an unknown environment, learning through a combination of exploration and exploiting acquired knowledge to maximize a long-term reward [1,2]. In [3], a model-based RL technique was proposed to optimize adaptive transmission and maximize EE. In [4], an adaptive power allocation based on Dyna-Q was proposed to minimize the energy consumption of the FSK communication system in time-varying underwater channels. A MAC protocol based on lightweight Q-learning was proposed in [5] to improve the EE of wireless sensor networks. In [6], a hot-booting Q-learning algorithm was proposed for underwater adaptive modulation and coding, incorporating an additional virtual learning stage to optimize the system performance including energy consumption. An improved RL method is proposed in [7] to maximize both energy efficiency and spectral efficiency for a UWA communication system. In [8], a Dyna-Q method is proposed to maximize the defined reward function of an underwater image communication system, and energy consumption plays a key role in evaluating its long-term performance.

Due to the use of neural networks in the RL methods with a continuous action space, there is always a higher probability of failure to converge [9]. Hence, the RL method with a discrete action space is more widely applied in UWA communication systems which have a relatively low-frequency interaction with the environment due to the low sound speed in the water and therefore have a stronger demand for reliability.

Existing works for EE maximization based on RL with discrete action spaces, such as Q-learning and Dyna-Q, use fixed action sets. However, prior knowledge is essential for the discretization of the action space like the transmission power. A small-size action space may fail to include the optimal transmission power, while the action space with a large size tends to increase the computation complexity and deteriorate its convergence performance. Moreover, there is an

[☆] This work was supported by the National Science Foundation grant ECCS-1651135 (CAREER).

[*] Corresponding author.
E-mail address: fancheng@mtu.edu (C. Fan).

https://doi.org/10.1016/j.phycom.2023.102218
Received 22 August 2023; Received in revised form 14 October 2023; Accepted 26 October 2023
Available online 30 October 2023

expected applicability degradation to the fixed action space of RL in a time-varying environment.

As per the use of a model, RL algorithms can be categorized into two types: model-based methods and model-free methods. Dyna-Q is a specific RL algorithm that combines both approaches to learn optimal policies in stochastic environments. Its effectiveness has been demonstrated in various domains, including power control in wireless communication systems. It can be leveraged to learn the optimal power control policy. In the mentioned approach, an RL agent is employed at the transmitter, which interacts with the environment to determine the most suitable power level for transmission. The agent receives rewards based on the performance metric defined based on EE and subsequently updates its policy accordingly.

In this work, an adaptive action space is introduced in the proposed RL method. Both simulations and experimental data processing reveal a superior performance of the proposed method compared with the other two comparative methods.

The main contribution of this work is summarized as follows.

- An action space with an adaptive size is proposed in the Dyna-Q algorithm. Besides the removal of the prior knowledge requirement, it allows the agent to explore and add potential optimal actions to its action space to improve the performance;
- Additionally, a cumulative modified EE is proposed to evaluate the system performance. Instead of summing the EE of all the transmission blocks directly, the energy consumption of failed communication is also taken into account;
- and a field trial was conducted to evaluate the performance of the proposed method along with comparative methods.

The rest of the paper is structured as follows. Section 2 presents the formulation of the optimization problem in the RL learning framework for adaptive switching among communication transmission power levels. Section 3 introduces the Dyna-Q method and an adaptive action space to address the optimization problem. The convergence of the proposed method is demonstrated in Section 4 using simulations in a measured underwater acoustic channel. Section 5 presents the experimental data processing and results. Finally, Section 6 draws the conclusion.

2. Problem formulation

2.1. System model

This work considers point-to-point underwater acoustic transmissions in blocks and assumes that the UWA channel is quasi-stationary for each block. To maximize long-term EE, the system can transmit the communication signal with an adaptive power level while ensuring its communication quality.

The underwater acoustic communication power control problem can be formulated as an MDP in which the transmitter's action is to choose an appropriate power level. The system's state is a function of the transmitted power and channel conditions. The MDP is defined as $M = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ with

- a discrete state space $\mathcal{S}$ specifying the system or the channel condition;
- a discrete action space $\mathcal{A} = \left\{a_1, a_2, \dots, a_M\right\}$ where the discrete action $a_m, m = 1, \dots, M$, specifies the mth communication strategy;
- the state transition probability P;
- the reward R when the state transits to another with an action from $\mathcal{A}$; and
- a discount factor $\gamma \in (0, 1]$.

The cumulative reward with the discount factor can be maximized to optimize the long-term system performance,

$$R_{\max} = \max V(s) = \max_{l=1,2,\dots,L} \mathrm{E}\left(\sum_{l=1}^{L} \gamma^l R^l\right), \tag{1}$$

where L is the total number of transmissions, and $\mathrm{E}(\cdot)$ is the expectation operator of a random variable.

Those elements in the MDP are described next in more detail.

2.1.1. The discrete state space

The state of the UWA communication power control system is defined as the received signal-to-noise ratio (SNR). It is one of the most widely-used indicators for the relationship between the received signal power and the noise level and can be measured simply at the receiver side [10–12].

$$\mathrm{SNR} := 10\log_{10}\left[\frac{P_{\mathrm{rx}} - P_{\mathrm{noise}}}{P_{\mathrm{noise}}}\right], \tag{2}$$

where P_{rx} is the received signal power and P_{noise} is the noise power.

2.1.2. The discrete action space

Communication signals with multiple power levels are implemented in the action space $\mathcal{A} = \left\{a_1, a_2, \dots, a_M\right\}$ and the action $a_m (m = 1, \dots, M)$ specifies the communication signal with mth transmission power level.

2.1.3. The reward function

Define N_{bits} as the number of information bits in each transmission block. Define T_{block} as the block time duration, and define B as the system bandwidth. Define P_{block} as the transmission power. Define p_{s} as the average bit success rate in each block transmission. To evaluate the performance of the system, the EE can be defined as

$$E_{\mathrm{e}} := \begin{cases} (N_{\mathrm{bit}} \times p_{\mathrm{s}})/(P_{\mathrm{block}} \times T_{\mathrm{block}}), & p_{\mathrm{s}} \geq p_{\mathrm{t}} \\ 0, & p_{\mathrm{s}} < p_{\mathrm{t}} \end{cases} \tag{3}$$

where p_{t} is a target average bit success rate. In practical systems, the transmission block with a bit success rate lower than p_{t} would be considered as a failed transmission with $E_{\mathrm{e}} = 0$. To maximize the long-term EE, the reward function is defined as

$$R := \begin{cases} E_{\mathrm{e}}, & p_{\mathrm{s}} \geq p_{\mathrm{t}} \\ -10, & p_{\mathrm{s}} < p_{\mathrm{t}} \end{cases} \tag{4}$$

where the negative number -10 is a penalty to accelerate converging by preventing the agent from executing certain actions.

2.2. Optimization

The Bellman optimality equation (BOE) is a commonly employed concept in RL due to the impracticality of directly solving Eq. (1) because of the large number of possible actions and states [13]. The BOE determines the best action to take in each state, considering both the immediate reward and the expected future discounted rewards.

Since the state transition function in the MDP considered here is not available, the Q-function, denoted as $Q(s^l, a^l)$, represents the expected return when taking action a^l in state s^l. It is defined as the expected value of the sum of the immediate reward R and the maximal Q-value in the next state s^{l+1} discounted by a factor of γ, namely,

$$Q(s^l, a^l) = \mathrm{E}[R + \gamma \max_{a^{l+1} \in \mathcal{A}} Q(s^{l+1}, a^{l+1})]. \tag{5}$$

The optimal action a^l_{opt} for the lth state, which maximizes the expected cumulative reward, can be determined by finding the argument that maximizes the Q-value,

$$a^l_{\mathrm{opt}} = \arg\max_{a \in \mathcal{A}} Q(s^l, a^l). \tag{6}$$

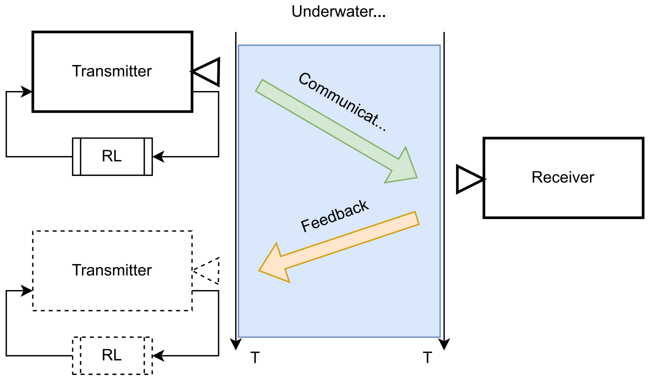


Fig. 1. Diagram of the RL-assisted UWA communication system.

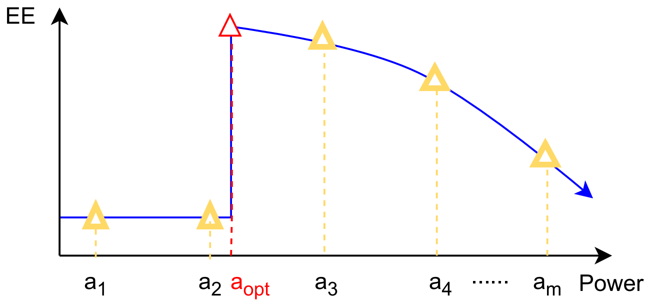


Fig. 2. Diagram of the potential optimal action searching.

3. The proposed method

3.1. Dyna-Q algorithm

Based on the channel condition, a Dyna-Q algorithm with an adaptive state space and an adaptive action space is employed for transmission power control in UWA communications.

The Dyna-Q is an RL algorithm. It updates the Q-value using both real experiences from the environment and simulated experiences generated by a model to solve the RL problems. It combines both model-free learning and model-based learning by incorporating a Q-learning update step along with a model of the environment to improve learning performance [14].

In the direct RL component, sequences of states, actions, and rewards obtained from real interactions with the environment are required. Based on the Eq. (5), there is an iterative form to approximate the expectation,

$$Q(s^l, a^l) \leftarrow (1-\alpha)Q(s^l, a^l) + \alpha[R + \gamma \max(Q(s^{l+1}, a^{l+1}))]. \tag{7}$$

where $\alpha \in (0, 1)$ is the learning rate.

The experiences from the real interactions are used for the model-learning component via $M(s^l, a^l) \leftarrow (R, s^{l+1})$, where M is the model to store the relationship between the tuple (s^l, a^l) and the tuple (R, s^{l+1}). With an estimated environment model, it becomes possible to determine the subsequent state and reward by utilizing the present state and action. By considering all potential future states, rewards, and their corresponding probabilities, the Q-value function can be computed within the framework of a model-based approach,

$$\begin{aligned} Q(s^l, a^l) \leftarrow \sum_{a^l} \pi(a^l|s^l) \sum_{s^{l+1},R} T(s^{l+1}, R|s^l, a^l)[R \\ + \gamma \max(Q(s^{l+1}, a^{l+1}))], \end{aligned} \tag{8}$$

where $T(s^{l+1}, R|s^l, a^l)$ is the model's estimate of the probability of transitioning from state s^l to state s^{l+1} when taking action a^l, and $\pi(a^l|s^l)$ is the probability of taking action a^l in the given state s^l following policy π.

3.2. Intelligent switching based on Dyna-Q with an adaptive action space

For an RL-assisted underwater acoustic communication system in Fig. 1, the receiver will send a feedback signal with necessary information once the processing for the communication signal is completed. Based on the feedback information, the transmitter is expected to choose optimal actions such as the power levels in this paper via RL algorithms.

Based on the definition of EE, it is clear that Eq. (3) is a piecewise function determined by the target average bit success rate $p_t = 1 - p_e$. A commonly used target bit error rate p_e is in the range of 10^{-2} to 10^{-6} depending on the system's objectives, the desired reliability, and the trade-off between the achievable data rate and the acceptable error rate. Hence, with the assumption of $p_e \ll 1$ when $p_s \geq p_t$, Eq. (3) can be rewritten as,

$$E_e \approx N_{\text{bit}}/(P_{\text{block}} \times T_{\text{block}}). \tag{9}$$

The EE is determined by the transmission power when $p_s \geq p_t$, since N_{bit} and T_{block} are fixed for all the transmission signals considered in this work.

Discretization is an essential part of the RL algorithms with discrete spaces like Dyna-Q. In this work, prior knowledge is required to determine a decent step size within a certain transmission power range to make the designed optimal action accessible to the agent. Moreover, in the time-varying UWA environment, a perfect design of the discrete space may not be ideal in future time blocks. In this work, an adaptive action space instead of a fixed equally divided action space is proposed for improved performance in EE. If all the transmission power levels are ordered from the lowest to the highest in the action space, an example of the discrete actions in a fixed action space and of the potential optimal action are illustrated in Fig. 2. For the action set $A_0 = \{a_1, a_2, a_3, a_4, \dots a_m\}$ in the fixed action space shown in Fig. 2, the equally divided transmission power levels yield different EE, and a_3 leads to the highest EE. However, the optimal action a_{opt} with a higher EE than a_3 is not included in the action space and is not accessible to the agent. It is clear that a search for the optimal action should be continued after reaching the convergence of a_3 to obtain a better EE performance.

Here, a trigger function is defined below to start the search for the potential optimal action,

$$L_a = \frac{L_0}{[N_a/N_t]}, \tag{10}$$

where L_a is the step length of the action searching; L_0 is the initial step length of an equally-divided action space in the design stage; $[\cdot]$ is the integer division operator; N_a is the number of continuous selections of a certain action; and N_t is the target number of continuous selections which determines the tendency for potential optimal action searching. Once a different action is executed, N_a will be reset as 1.

Take Fig. 2 as an example. Once action a_3 has been executed N_a times, it will be considered as the optimal one in the current action space, and the potential optimal action searching will be triggered. Based on the basis of binary search, the length of the searching step on the left is slashed in half and a temporal action a_{temp} is added to the action space at that time until the minimal step length is reached. Its initial Q-value is copied from a_3 to encourage the exploration.

The Dyna-Q algorithm with adaptive action space is summarized in Algorithm 1.

4. Simulations

There are various techniques for underwater acoustic communications [15,16]. In this work, the orthogonal frequency-division multiplexing (OFDM) technique is employed for simulation and experimental

Algorithm 1 Dyna-Q Algorithm with adaptive action space

```
Initialize Q-table: $Q(s^l, a^l)$ for all state–action pairs
Initialize Model: $M(s^l, a^l)$ for all state–action pairs with default values
Initialize environment state: $s$
while episode $\leq T_e$ do
    Choose an action $a^l$ via epsilon-greedy based on Q-table
    Take action $a^l$ and observe the reward $R$ and the next state $s^{l+1}$
    if $a^l = a^{l-1} = a_m$ then
        $N_a \leftarrow N_a + 1$
    else
        $N_a \leftarrow 1$
    end if
    if $N_a \geq N_t$ then
        $\{a_1, a_2, \dots, a_{m-1}, a_{m'}, a_m, \dots, a_M\} \leftarrow \{a_1, a_2, \dots, a_{m-1}, a_m, \dots, a_M\}$ $(m-1 < m' < m)$
        $Q(s^l, a^l_{m'}) \leftarrow Q(s^l, a^l_m)$
    end if
    Update Q-table: $Q(s^l, a^l) = (1-\alpha)Q(s^l, a^l) + \alpha[R + \gamma \max(Q(s^{l+1}, a^{l+1}))]$
    Update the model: $M(s^l, a^l) \leftarrow (R, s^{l+1})$
    while steps $\leq T_s$ do
        Sample a random state–action pair $(s^l, a^l)$ from the model
        Retrieve the predicted reward and next state from the model: $(R, s^{l+1}) \leftarrow M(s^l, a^l)$
        Update Q-table using the model: $Q(s^l, a^l) = (1-\alpha)Q(s^l, a^l) + \alpha[R + \gamma \max(Q(s^{l+1}, a^{l+1}))]$
    end while
    Set the current state to the next state: $s^l \leftarrow s^{l+1}$
end while
```

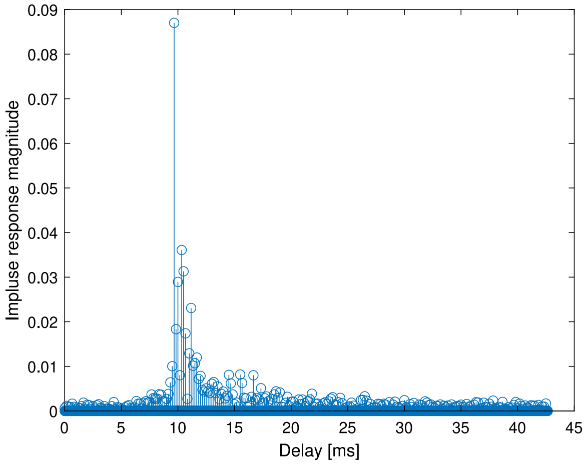


Fig. 3. The impulse response of KW channel.

data processing. The OFDM is widely used due to its high spectral efficiency and robustness to multipath fading [17,18].

The proposed method is first evaluated in a measured underwater acoustic channel impulse response (CIR) shown in Fig. 3 with additive white Gaussian noise (AWGN). The CIR was measured on 2022-08-10 at 21:33:48 UTC in the Keweenaw Waterway (KW), Houghton, Michigan, USA.

4.1. Signaling method

The OFDM structure for simulation is shown in Fig. 4. The waveform includes multiple preambles, 20 zero-padding (ZP)-OFDM data blocks, 1 s for recording background noise, and a Hyperbolic Frequency Modulation (HFM) chirp as postamble. Linear Frequency Modulated (LFM) Pulse Waveform is used for synchronization. The OFDM data blocks consist of 1024 subcarriers, uniformly distributed in a 6 kHz bandwidth. These subcarriers consist of 256 pilot subcarriers, 672 data subcarriers, and 96 null subcarriers. The lowest and highest 32 subcarriers are null subcarriers, while one pilot subcarrier is present in every 4 adjacent subcarriers. The data subcarriers are situated in the middle-frequency range, and there is one null subcarrier in every 20 adjacent subcarriers. Each data block has a duration of 170 ms and is followed by an 80 ms guard time. The frequency range used for operations in all communication strategies is between 21 kHz and 27 kHz. Table 1 provides detailed information about the associated parameters and processing methods for OFDM communication.

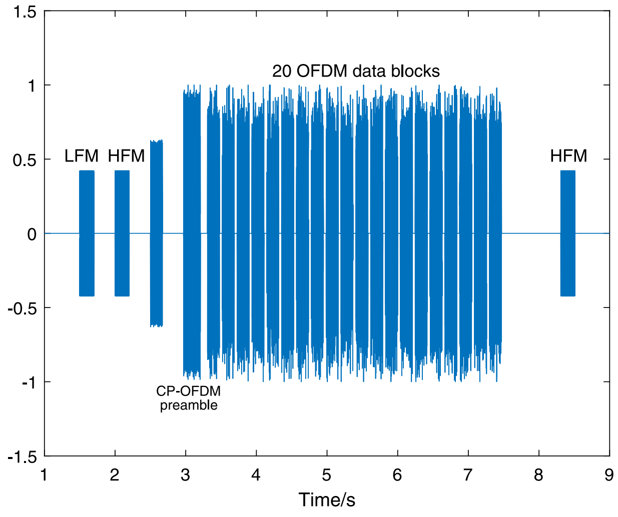


Fig. 4. The transmitted OFDM waveform.

Table 1
Parameters of the OFDM communication.

Modulation	QPSK
Bandwidth (Hz)	6000
Baud rate (bits/s)	7875
Channel estimation	Least Square
Equalization	Least Square
Error correction	×

4.2. Simulation results

In each Monte Carlo simulation, the transmitted OFDM signal passes through the KW channel. White Gaussian noise of different levels is added to the channel output to simulate different SNR scenarios.

At the receiver side, the SNR of the received signal is estimated as follows. The guard interval before the postamble is used to record the noise and estimate the noise power. The power of the useful signal plus noise is estimated based on the OFDM data blocks. Then the SNR can be calculated based on Eq. (2). In practice, the estimation of the received SNR is utilized in this approach as the channel state, which makes it necessary to include estimation error in the simulation to evaluate its performance improvement.

Based on the decoding outcomes from 1000 data blocks, the BER under different estimated SNR conditions in the KW channel is depicted in Fig. 5. Although the resulting curve may not be a perfect waterfall due to the SNR estimation error, it enhances the efficacy of the simulation as compared to employing theoretically computed received SNR values.

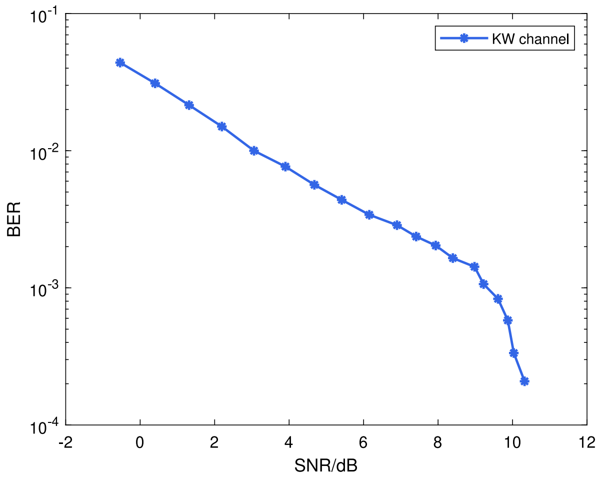


Fig. 5. BER of OFDM transmission in the KW channel.

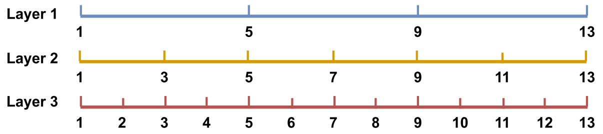


Fig. 6. The action space with different layers. The X-axis is the transmission power level in dB.

Table 2
The state space.

State	SNR
s_1	SNR < 7 dB
s_2	SNR < 11 dB
s_3	SNR < 15 dB
s_4	SNR ≥ 15 dB

4.2.1. Testing schemes

Based on the simulated BER curve, we define a state space S_4 as listed in Table 2. To examine the improvement of the proposed method with an adaptive action space, a three-layer action space within the same transmission power range is defined in Fig. 6. Specifically, there are 4 discrete power levels 1, 5, 9, and 13 in Layer 1., By halving the transmission power difference, power levels 3, 7, and 11 are added to Layer 2. Similarly, more power levels are added to Layer 3 to have smaller power steps within the same transmission power range.

We consider two comparative schemes:

- Scheme 1: The Dyna-Q algorithm with a state space of S_4, and a fixed action space of Layer 1; and
- Scheme 2: The Dyna-Q algorithm with a state space of S_4, and a fixed action space of Layer 3.

where the comparative Dyna-Q algorithms with fixed action space are summarized in Algorithm 2.

The proposed method initializes its adaptive action space using Layer 1 and then adds the potential optimal action of the next two layers based on the convergence.

A discount factor $\gamma = 0.9$ and a number of planning steps of 100 are used in the three methods. The other hyper-parameters are also the same for the three methods.

4.2.2. The cumulative EE

To evaluate the long-term EE performance of a point-to-point UWA communication system, we introduce the cumulative EE C_e as the sum of the EE of all the data blocks,

$$C_e := \sum_{i=1}^{I} E_{\mathrm{e}}^{i}, \tag{11}$$

where E_{e}^{i} is the EE of the ith block, and I refers to the total number of data blocks.

Algorithm 2 Dyna-Q Algorithm with fixed action space

```
Initialize Q-table: $Q(s^l, a^l)$ for all state–action pairs
Initialize Model: $M(s^l, a^l)$ for all state–action pairs with default values
Initialize environment state: $s$
Initialize action space with the given layer: $A_L$
while episode $\le T_e$ do
    Choose an action $a^l \in A_L$ via epsilon-greedy based on Q-table
    Take action $a^l$ and observe the reward $R$ and the next state $s^{l+1}$
    Update Q-table: $Q(s^l, a^l) = (1 - \alpha)Q(s^l, a^l) + \alpha[R + \gamma \max(Q(s^{l+1}, a^{l+1}))]$
    Update the model: $M(s^l, a^l) \leftarrow (R, s^{l+1})$
    while steps $\le T_s$ do
        Sample a random state–action pair $(s^l, a^l)$ from the model $M$
        Retrieve the predicted reward and next state from the model: $(R, s^{l+1}) \leftarrow M(s^l, a^l)$
        Update Q-table using the model: $Q(s^l, a^l) = (1 - \alpha)Q(s^l, a^l) + \alpha[R + \gamma \max(Q(s^{l+1}, a^{l+1}))]$
    end while
    Set the current state to the next state: $s^l \leftarrow s^{l+1}$
end while
```

4.2.3. The cumulative modified EE

For the failed communication attempts that have smaller bit success rates than the defined target rate, the energy efficiencies are considered zero based on the definition, regardless of the corresponding transmission power. Here, we propose a cumulative modified EE which takes the wasted energy of the failed communication into account to evaluate the long-term performance. Rather than directly summing up the EE of all the data blocks in the cumulative EE, the cumulative modified EE is defined as

$$C_{me} := \frac{\sum_{i=1}^{I} N_{\mathrm{bits}}^{i}}{\sum_{i=1}^{I} (P_{\mathrm{block}}^{i} \times T_{\mathrm{block}})}, \tag{12}$$

where the number of cumulative bits for the ith block N_{bits}^{i} is defined as

$$N_{\mathrm{bits}}^{i} := \begin{cases} N_{\mathrm{bit}} \times p_{\mathrm{s}}, & p_{\mathrm{s}} \ge p_{\mathrm{t}} \\ 0, & p_{\mathrm{s}} < p_{\mathrm{t}}. \end{cases} \tag{13}$$

4.2.4. Simulation results

Assuming perfect feedback from the receiver, Fig. 7 shows the cumulative EE of the three schemes. One can see that the proposed method has the best performance. The mean of the EE for the proposed method is 4253.9 bps/W, while that for Scheme 1 and Scheme 2 are 2788.8 bps/W and 4166.2 bps/W respectively.

Fig. 8 depicts the cumulative modified EE of the three schemes. One can see that the proposed method has a better performance than the two comparative schemes. For the proposed method, due to the small size of the initial action space, the cumulative modified EE grows as fast as Scheme 1 which has an action space of size 4. Then Scheme 1 converges to the optimal action within its action space, while the proposed method keeps searching for the potential optimal actions and adds them into its adaptive action space. The proposed method and Scheme 2 converge to the same level of cumulative modified EE. The cumulative modified EE is higher than that of Scheme 1 because the optimal action is chosen from a larger action space with a size

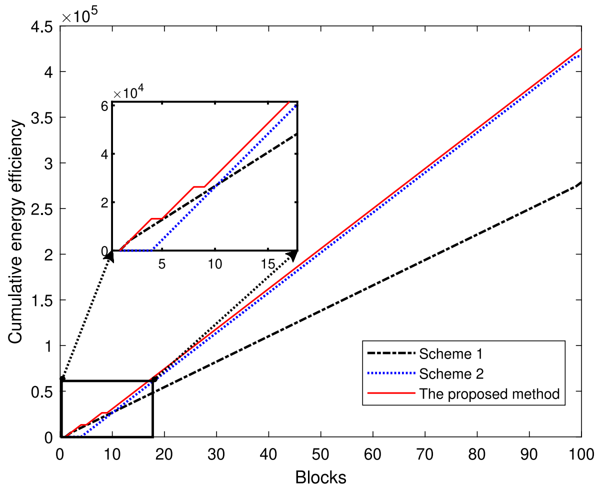


Fig. 7. Cumulative EE in the KW channel.

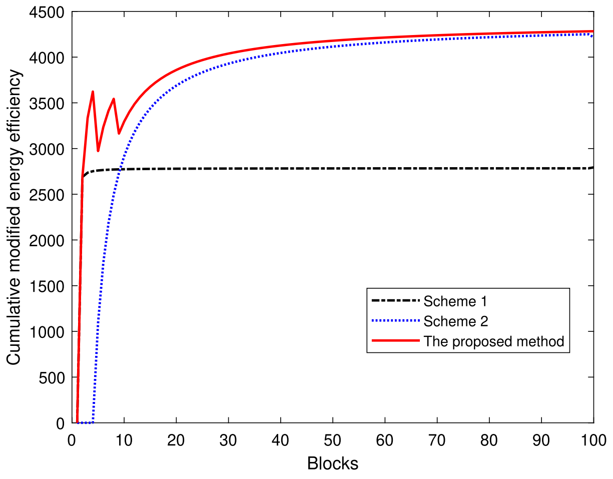


Fig. 8. Cumulative modified EE in the KW channel.

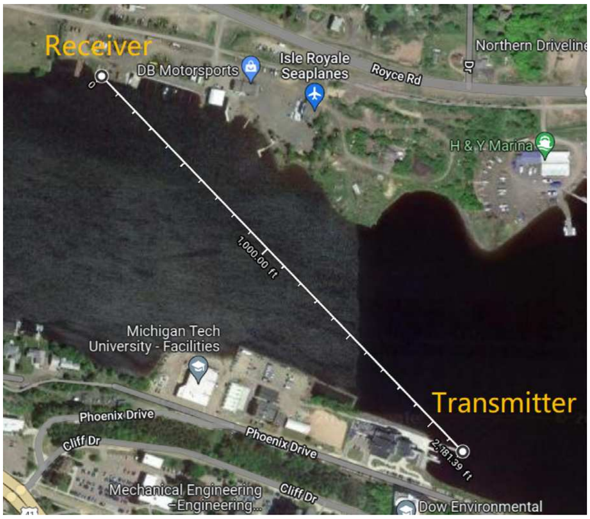


Fig. 9. Deployment of the transmitter and the receiver.

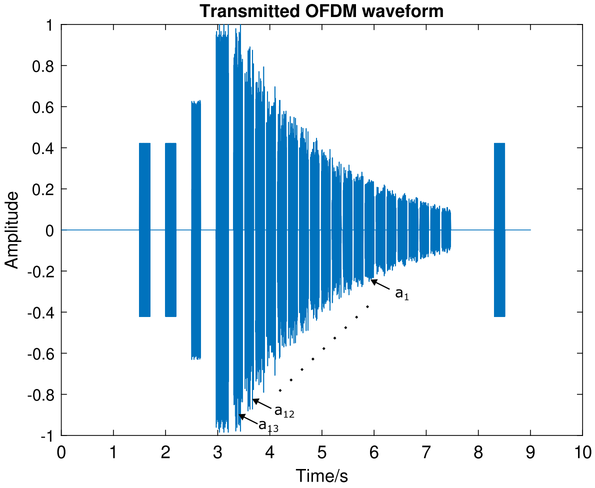


Fig. 10. The transmission waveform of an OFDM packet.

of 13. Moreover, although the proposed method and Scheme 2 have almost the same performance after convergence, the proposed method converges much faster than Scheme 2 and is expected to have a better performance in more frequently changing channels.

5. Experimental data processing

5.1. Experimental data collection

The experiment was conducted in April 2023 in the Keweenaw Waterway, Houghton, Michigan, USA. Fig. 9 depicts the layout of the transmitter and the receiver. The distance between them is about 675 m, and the depth of the transmitter is 2.5 m while that of the receiver is 1 m. The depth of the waterway nearby ranges from 2 m to 8 m during this experiment. Both the transmitter and the receiver drifted with waves. The receiving modem has four hydrophones with 10 cm spacing. This experiment is from 2023-04-28 at 23:04:40 UTC to 2023-04-30 at 15:48:36 UTC.

The transmission waveform of an OFDM packet is shown in Fig. 10. The waveform that is sent includes multiple preambles followed by 20 ZP-OFDM data blocks. The transmission power of the ZP-OFDM data block decreases by 1 dB with each block in the packet. Due to the space–time variety characteristic of the underwater acoustic channel, it is unfeasible to compare the real-time system performance with different RL algorithms in the same underwater environment. Assuming that the channel condition remains unchanged within each OFDM packet, through this design of the received waveforms with different transmission power levels (i.e., actions) are recorded. This allows the evaluation of different schemes based on the experimental data.

The total duration of the waveform for each packet is approximately 9 s. When it is transmitted by a commercial OFDM modem, an additional preamble is added for detection and synchronization. If this additional preamble is not detected or properly decoded, the receiving modem will not record the waveform and report a lost packet. The waveform is transmitted within the 21–27 kHz frequency range. The OFDM packet with the same structure is sent in every 60-second time slot, which is shown in Fig. 11.

5.2. Experimental data processing results

During the experiment, 543 packets out of 2527 packets were successfully recorded, while other lost packets are represented as empty spaces in Fig. 13. It shows the executed actions (i.e., transmission power levels) and the corresponding received SNRs for the 543 packets.

The first 13 transmission power levels (i.e., action) of each OFDM packet are included in the action space. For each valid packet, the action that yields the highest EE is highlighted in Fig. 12.

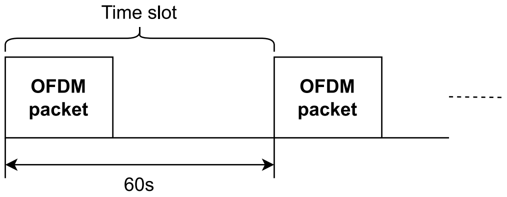


Fig. 11. The time-slotted OFDM transmission.

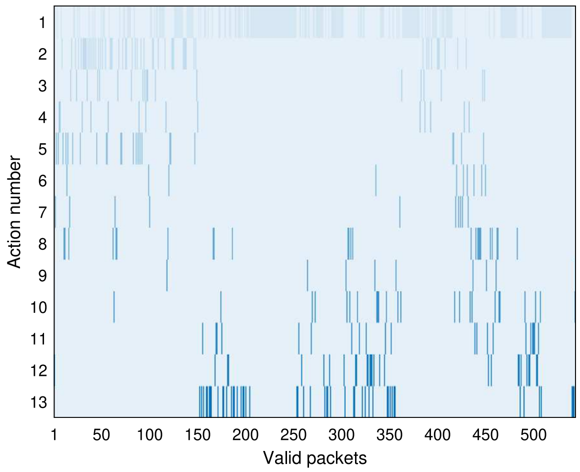


Fig. 12. The actions with the highest EE for all the valid packets.

It is evident that the optimal action varies according to the temporal fluctuation of the channel. Each action in the action space has been selected as an optimal action at least once.

To evaluate the performance of the proposed method and two comparative schemes, the performance metrics used in the simulation (the cumulative EE, and the cumulative modified EE) are employed in the experimental data processing.

The target average bit success rate keeps $p_s = 0.99$ since there is no error correction employed. There are 100 steps for planning within each time slot for the Dyna-Q algorithm in all methods. Namely, in each time slot, the agent keeps training 100 times based on the data from the previous attempts to find the optimal or near-optimal action. Other RL parameters like learning rate $\alpha = 0.1$ and discount factor $\gamma = 0.9$ keep the same for all the schemes as well.

5.2.1. Analysis based on executed actions

Fig. 13 depicts the EE of each action executed by the proposed method and two comparative schemes. Due to the small size of the initial action space in the proposed method, one can observe from the first 100 packets that the proposed method inherits the fast-boot characteristic from Scheme 1 which has smaller state space and action space and is expected to converge with less time consumption. Meanwhile, there are more exploration attempts compared to those in Scheme 1, and the proposed method before Scheme 2 converges to the optimal action, which shows the executed actions for Scheme 2 with smaller EE in that period.

There is a severe change in the channel condition around the 600th time slot. The actions executed by all the schemes yield small EE from the 600th packet to the 1500th packet. Once it reaches the stage between the 1600th and 1900th packet where the channel conditions are relatively good and stable, all schemes exhibit an increase in EE. While the proposed method executes more suboptimal actions compared to Scheme 1 due to its adaptive action space with temporal actions added, the performance gap is notably smaller than that observed between Scheme 2 and the others. The temporal actions show the benefit when the SNR declines in the subsequent 600 packets. Unlike Scheme 1 which directly switches from actions with EE over 4000 bps/W to actions with EE of approximately 2000 bps/W, the proposed method allows the selection of several temporal actions to adapt to the changing channel condition and yields a better overall performance compared to the other two schemes during this period. From Table 3, one can see that the proposed method has the highest mean EE of 1771.4 bps/W, while that for Scheme 1 and Scheme 2 are 1510.4 bps/W and 1504.9 bps/W, respectively.

Table 3
EE of different schemes.

Scheme	Average of the modified EE
The proposed method	1771.4 bps/W
Scheme 1	1510.4 bps/W
Scheme 2	1504.9 bps/W

5.2.2. Performance analysis based on cumulative EE and cumulative modified EE

The long-term EE performance analysis is conducted based on 543 packets that were successfully recorded in the experiment. Fig. 14 shows the cumulative EE of the proposed method and two comparative schemes.

In the first 120 packets, the cumulative EE of the proposed method grows more rapidly than the other two schemes. From the 120th to the 320th packet, all the schemes have a negligible increase in the cumulative EE until the channel condition turns better. Around the 320th packet, the proposed method exhibits a timely response and yields the best performance.

Fig. 15 depicts the cumulative modified EE of three different methods. One can see that the proposed method still has the best performance. Due to the small size of the initial action space in the proposed method, the cumulative modified EE grows rapidly in the first 10 valid packets. For Scheme 2, due to its large action space, it spends an extra exploration cost in the first 40 valid packets before it converges to the optimal action, and then exhibits similar performance with the proposed method from the 40th packet to the 300th packet. Scheme 1 has the same rapid-growing performance as the proposed method in the first 10 valid packets, but from the 10th to the 120th packet, it fails to execute the same optimal action as the other two schemes due to its limited action space. When the channel condition becomes better around the 320th packet, the cumulative modified EE of all the schemes increases again, and the proposed method still has the best performance.

5.2.3. Computational complexity analysis

Denote $|S|$ as the number of states, $|A|$ as the number of actions, and N_p as the number of planning steps in Dyna architecture, the computational complexity of Dyna-Q is $O(N_p|S|ln|A|)$ [19,20]. Note that the proposed method utilizes the same state space and number of planning steps as the comparative methods, only the sizes of action spaces determine the computational complexity rank. Compared with Scheme 2, the proposed method has a smaller action space, and therefore a lower computational complexity to converge to the same optimal action of Scheme 2. Due to the designed action space adaptation in the proposed method, only the potential optimal actions will be added to the action space gradually. The extra computational complexity of the proposed method compared with Scheme 1 is considered moderate and acceptable.

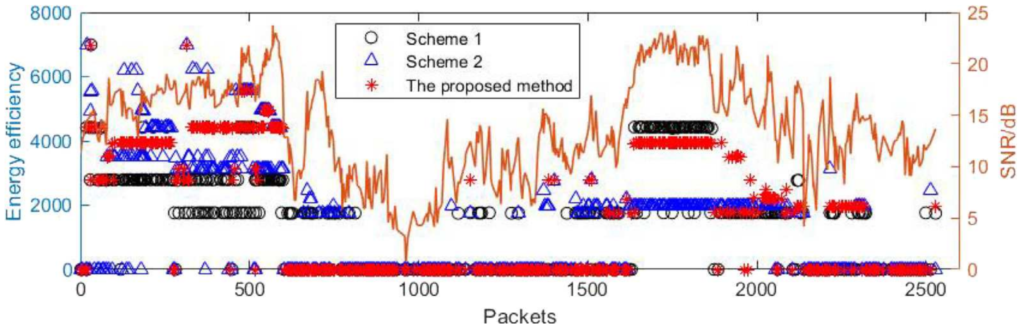


Fig. 13. Reward of actions executed in three schemes.

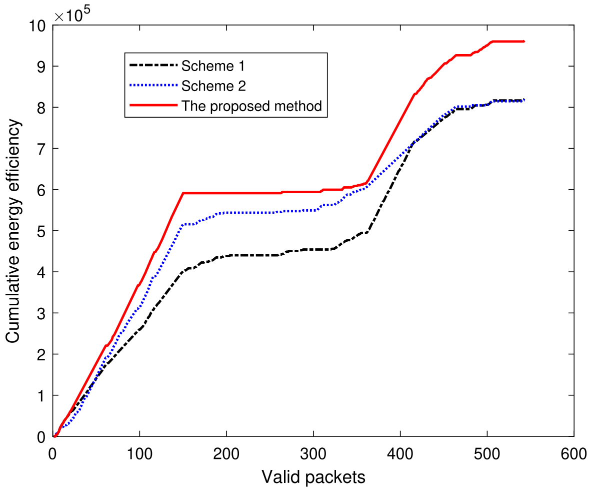


Fig. 14. Cumulative EE of three schemes in the experimental data processing.

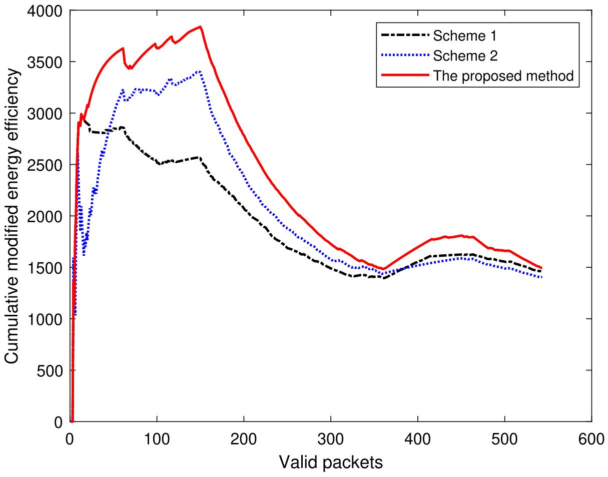


Fig. 15. Cumulative modified EE of three schemes in the experimental data processing.

6. Conclusions

This work studies adaptive power control for UWA communication systems. A Dyna-Q algorithm with an adaptive action space was proposed to find the optimal transmission power level for each time slot, to maximize the overall EE performance of the system. Simulations in the measured underwater acoustic channel examined its converging behavior and showed that the proposed method achieves better performance than the two comparative methods. Its superior performance in fast converging and energy efficiency was also examined and demonstrated in the field of experimental data processing.

7. Future research

The proposed method with adaptive action space is developed to search for the optimal action more efficiently based on the system model without suboptimal. However, future research is required to improve the optimal searching among multiple suboptimal.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Cheng Fan reports financial support was provided by National Science Foundation.

Data availability

Data will be made available on request.

References

[1] R. Sutton, A. Barto, Reinforcement Learning: An Introduction, MIT Press, Massachusetts, 2018.

[2] P. Dayan, Reinforcement learning, in: Stevens' Handbook of Experimental Psychology, Vol. 3, No. 5, 2002, pp. 103–129.

[3] C. Wang, Z. Wang, W. Sun, D. Fuhrmann, Reinforcement learning-based adaptive transmission in time-varying underwater acoustic channels, IEEE Access 6 (5) (2017) 2541–2558, http://dx.doi.org/10.1109/ACCESS.2017.2784239.

[4] Y. Zhao, L. Wan, E. Cheng, F. Xu, Adaptive power allocation for non-coherent FSK in time-varying underwater acoustic communication channels, in: 2022 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC, IEEE, 2022, pp. 1–6.

[5] C. Savaglio, P. Pace, G. Aloi, A. Liotta, G. Fortino, Lightweight reinforcement learning for energy efficient communications in wireless sensor networks, IEEE Access 7 (2019) 29355–29364.

[6] W. Su, J. Lin, K. Chen, L. Xiao, C. En, Reinforcement learning-based adaptive modulation and coding for efficient underwater communications, IEEE Access 7 (5) (2019) 67539–67550, http://dx.doi.org/10.1109/ACCESS.2019.2918506.

[7] C. Fan, L. Wei, Z. Wang, Adaptive switching for communication profiles in underwater acoustic modems based on reinforcement learning, Appl. Acoust. 210 (2023) 109430, http://dx.doi.org/10.1016/j.apacoust.2023.109430.

[8] W. Su, J. Tao, Y. Pei, X. You, L. Xiao, E. Cheng, Reinforcement learning based efficient underwater image communication, IEEE Commun. Lett. 25 (3) (2021) 883–886, http://dx.doi.org/10.1109/LCOMM.2020.3041937.

[9] A. Larsson, Reinforcement learning in problems with continuous action spaces: a comparative study, 2021.

[10] Y. Chen, W. Yu, X. Sun, L. Wan, Y. Tao, X. Xu, Environment-aware communication channel quality prediction for underwater acoustic transmissions: A machine learning method, Appl. Acoust. 181 (2021) 108128.

[11] G. Song, X. Guo, W. Wang, Q. Ren, J. Li, L. Ma, A machine learning-based underwater noise classification method, Appl. Acoust. 184 (2021) 108333.

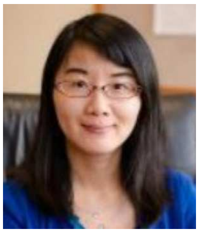
[12] Y. Qiu, F. Yuan, S. Ji, E. Cheng, Stochastic resonance with reinforcement learning for underwater acoustic communication signal, Appl. Acoust. 173 (2021) 107688.

[13] S. Dutta, Reinforcement Learning with TensorFlow, Packt Publishing, 2018.

[14] M. Pei, H. An, B. Liu, C. Wang, An improved dyna-q algorithm for mobile robot path planning in unknown dynamic environment, IEEE Trans. Syst. Man Cybern. Syst. 52 (7) (2021) 4415–4425.
[15] K.B. Yoo, G.F. Edelmann, Low complexity multipath and Doppler compensation for direct-sequence spread spectrum signals in underwater acoustic communication, Appl. Acoust. 180 (2021) 108094.
[16] X. Hu, D. Wang, Y. Lin, W. Su, Y. Xie, L. Liu, Multi-channel time frequency shift keying in underwater acoustic communication, Appl. Acoust. 103 (2016) 54–63.
[17] S. Zhou, Z. Wang, OFDM for Underwater Acoustic Communications, Wiley, Germany, 2014.
[18] R. Chen, W. Wu, Q. Zeng, S. Liu, Construction and application of polar codes in OFDM underwater acoustic communication, Appl. Acoust. 211 (2023) 109473.
[19] B.H. Abed-alguni, M.A. Ottom, Double delayed Q-learning, Int. J. Artif. Intell. 16 (2) (2018) 41–59.
[20] M. Maroto-Gómez, R. González, Á. Castro-González, M. Malfaz, M.Á. Salichs, Speeding-up action learning in a social robot with dyna-q+: A bioinspired probabilistic model approach, IEEE Access 9 (2021) 98381–98397.

Cheng Fan received the B.S. degree in environmental engineering from the Northwestern Polytechnical University, China, in 2016, and the M.S. degree in communication engineering from the Harbin Engineering University, China, in 2019.

He is currently a Research Assistant at the Michigan Technological University, Houghton, MI, where he is working towards the Ph.D. degree in electrical engineering. His research interests include underwater acoustic communications and reinforcement learning employed in underwater environments.



ZHAOHUI WANG (S'10–M'13) received the B.S. degree from the Beijing University of Chemical Technology in 2006, the M.S. degree from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2009, and the Ph.D. degree from the University of Connecticut, Storrs, in 2013, all in electrical engineering. She has been with the Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, as an Assistant Professor, since 2013. Her research interests lie in the areas of wireless communications, networking, and statistical signal processing, with a recent focus on signal processing and machine learning techniques for wireless communications and networking in underwater acoustic environments. Dr. Wang was honored with the Women of Innovation Award by the Connecticut Technology Council in 2013. She was a recipient of the NSF CAREER Award in 2017. She served as a technical reviewer for many premier journals and conferences. She was recognized as an Outstanding Reviewer by the IEEE JOURNAL OF OCEANIC ENGINEERING from 2012 to 2016.

Dr. Yang received his Ph.D. in Electrical Engineering from the University of Florida in 2022, under the supervision of Professor Shuo Wang, Professor Yier Jin, and Professor Yuguang Fang. Afterward, he joined the Department of Electrical and Computer Engineering as an assistant professor.

Dr. Yang's research interests lie in deep learning-related cybersecurity, hardware security, and network security. He published many papers and presented at top-tier AI and security conferences such as AAAI, AsiaCCS, NDSS, and DAC. He also served as a reviewer of several top journals and conferences.